REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
NOV 2015	CONFERENCE PAPER (Post Print)	May 2013 – Mar 2015

4. TITLE AND SUBTITLE

Memristive Computational Architecture of an Echo State Network for Real-Time Speech-Emotion Recognition

5a. CONTRACT NUMBER
IN-HOUSE

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62788F

6. AUTHOR(S)

Q. Saleh*, C. Merkel*, D. Kudithipudi*, B. Wysocki

5d. PROJECT NUMBER
SSNT

5e. TASK NUMBER
IN

5f. WORK UNIT NUMBER
HO

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

*NanoComputing Research Lab
Rochester Institute of Technology
Rochester, NY 14623

Air Force Research Laboratory /RITB
525 Brooks Road.
Rome, NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/Information Directorate
Rome Research Site/RITB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2015-006

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA Case Number: 88 ABW-2015-0832
DATE CLEARED: 5 MAR 2015

13. SUPPLEMENTARY NOTES
 Proceedings IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), Verona NY, 26-28 May 2015. ISBN: 978-1-4673-7556-6.

14. ABSTRACT

Echo state networks (ESNs) provide an efficient classification technique for spatiotemporal signals. The feedback connections in the ESN topology enable feature ex-traction of both spatial and temporal components in time series data. This property has been used in several application domains such as image and video analysis, anomaly detection, and speech recognition. In this research, a hardware architecture was explored for realizing ESN efficiently in power-constrained devices. Specifically, a scalable computational architecture applied to speech-emotion recognition was proposed. Two different topologies were explored, with memristive synapses. The simulation results are promising with a classification accuracy of ≈ 96% for two distinct emotion statuses.

15. SUBJECT TERMS
Memristor, Reservoir Computing, Echo State Networks, Speech Emotion Recognition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					CLARE D. THIEM
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (*Include area code*)
U	U	U	UU	6	N/A

Memristive Computational Architecture of an Echo State Network for Real-Time Speech-Emotion Recognition

Qutaiba Saleh, Cory Merkel, Dhireesha Kudithipudi
IEEE Student Member, IEEE Student Member, IEEE Senior Member
NanoComputing Research Lab
Rochester Institute of Technology
Rochester, NY 14623
Email: dxkeec@rit.edu

Bryant Wysocki IEEE Senior Member Information Directorate
Air Force Research Laboratory
Rome, NY 13441

***Abstract*—Echo state neural networks (ESNs) provide an efficient classification technique for spatiotemporal signals. The feedback connections in the ESN topology enable feature extraction of both spatial and temporal components in time series data. This property has been used in several application domains such as image and video analysis, anomaly detection, and speech recognition. In this research, we explore a hardware architecture for realizing ESN efficiently in power-constrained devices. Specifically, we propose a scalable computational architecture applied to speech-emotion recognition. Two different topologies are explored, with memristive synapses. The simulation results are promising with a classification accuracy of $\approx$ 96% for two distinct emotion statuses.**

***Keywords*—Reservoir Computing, Echo State Networks, Memristors, Speech Emotion Recognition**

I. Introduction

Echo State Network(ESN) is a class of reservoir computing model presented by Jaeger *et.al.* in 2001 [1]. ESNs are considered as partially-trained Artificial Neural Networks (ANNs) with a recurrent network topology. They are used for spatiotemporal signal processing problems, where signals are processed based on their behavior in time series windows. The ESN model is inspired by the emerging dynamics of how the brain handles temporal stimuli. It consists of an input layer, a reservoir layer and an output layer (see Fig. 1). The reservoir layer, is the heart of the network, with rich recurrent connections. These connections are randomly generated and each connection has a random weight associated with it. Once generated, these random weights are never changed during training or testing phases of the network. The output layer of the ESN linearly combines the desired output signal from the rich variety of excited reservoir layer signals. The central idea is that only the network to the output layer connection weights have to be trained, using simple linear regression algorithms. Another reservoir model, known as liquid state machine, provides a biologically plausible model for generating computations in cortical microcircuits.In contrast, ESN provides a high performance mathematical framework for solving a number of engineering tasks. Specifically, they can be applied to recurrent artificial neural networks without internal noise. ESNs have simplified training algorithms compared to other recurrent ANNs and are more efficient than kernel-based methods (e.g.: Support Vector Machines) due to their ability to incorporate temporal stimuli. Because of its recurrent connections, the output of the reservoir depends on the current input state and all previous input states within the system memory. The recurrent connections within the reservoir layer of the ESNs enable extracting both spatial and temporal components in the time series data. This attractive feature has been used in several applications that deal with spatiotemporal problems. Software implementations of ESN have been effective in diverse applications such as emotion recognition [2], forecasting of water inflow for a hydropower plant [3], natural language analysis [4], motion identification [5], speech recognition [6], and many more (see [7] for a review). However, the software models are not efficient for embedded and low-end processing environments, where power dissipation is critical to the operation of the devices. To address this, we propose a computational architecture of the ESN that enables portability and power efficiency by exploiting the use of emerging memristive devices. Specifically, the synapses are implemented using memristors. The neuron circuits used in this architecture are composed of current-mode designs,which are inherently low power. Two different ESN architecture topologies are explored with the memristive circuits, ring and random topology. The architecture is tuned and modeled for a speech-emotion recognition with 96% classification accuracy. Such models are proven to be efficient for several applications . For example, it has been demonstrated in auditory modeling that identifying and recognizing the emotional status of a person is crucial for designing effective next-generation human-computer interaction interface [2]. The rest of the paper is organized as follows. ESN training algorithm is discussed in section II. Section III provides an overview of the hardware architecture and the synapse circuit design. Section IV discusses speech emotion recognition application and the

The material and results presented in this paper have been cleared for public release, unlimited distribution by AFRL, case number 88ABW-2015-0832. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL or its contractors.

architecture. Section V presents the results for the proposed architecture and Section VI presents conclusions.

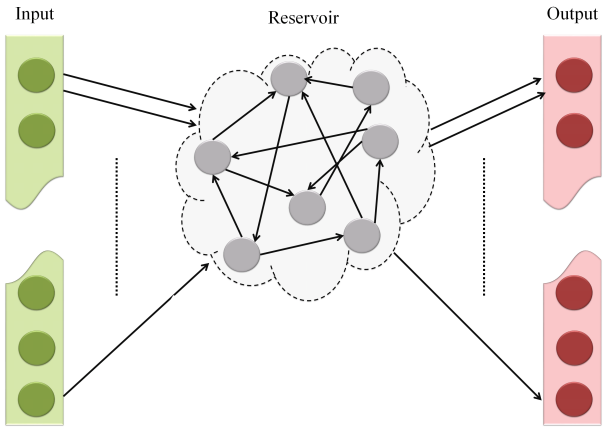

Fig. 1. Echo State Networks consist of three layers: input layer, reservoir layer and output layer.

II. ESN Training Algorithm

Three main sets of weights are associated with the ESNs (Fig. 2). The weights at the input and reservoir layers are randomly generated. These layers are used to extract temporal features of the input signal. They can be thought of as an internal pre-process step that prepares the signal for the actual processing layer where the classification is learned at the readout layer. Fig. 2 also shows the propagation of the signals through the ESNs. The input signal to the ESN $u(n)$ is pre-processed at the input and reservoir layers to extract the temporal featured signal $x(n)$ which is fed to the readout layer to complete the classification process. Considering that the input and reservoir layers are not actual parts of this process, their weights are not trained which makes training the ESNs much easier than other types of recurrent neural networks.

The goal of the training algorithm is to calculate the weights at the output layer based on the dynamic response(states) of the reservoir layer [8]. The states of the reservoir layer are calculated based on the input vectors and the weights of the input and reservoir layer as shown in (1).

$$\mathbf{x}(n+1) = f^{\text{res}}\left(\mathbf{W}_{\text{in}}\mathbf{u}(n+1) + \mathbf{W}_{\text{x}}\mathbf{x}(n)\right) \tag{1}$$

where $\mathbf{u}(n)$ is the ESN input, $\mathbf{W}_{\text{in}}$ is the weight matrix between the input layer and reservoir, $\mathbf{W}_{\text{x}}$ is the weight matrix between the neurons within the reservoir, and f^{res} is the reservoir's activation function. The states of the reservoir for all input vectors are used as an input to a supervised training to calculate the output weights $\mathbf{W}_{\text{out}}$. The normal equation is used to implement the supervised training of ESNs (2).

$$\mathbf{W}_{\text{out}} = (\mathbf{Y}.\mathbf{X}')(\mathbf{X}.\mathbf{Y}') \tag{2}$$

where $\mathbf{X}$ is a matrix concatenating all states of the reservoir and $\mathbf{Y}$ is a matrix of all training outputs.The process for training the ESN can be explained through the following steps:

1) At initialization, randomly generate the weights for the input and reservoir layers ($\mathbf{W}_{\text{in}}$ and $\mathbf{W}_{\text{x}}$)
2) Drive the next input vector $\mathbf{u}(n+1)$ to the input layer
3) Calculate the response of the reservoir layer by (1)
4) Save the response in a matrix ($\mathbf{X}$)
5) Repeat step 2-4 for all input vectors
6) Calculate output weights based on normal equation (2)

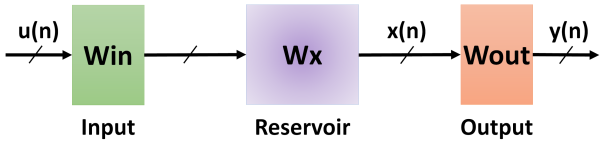

Fig. 2. Echo State Networks abstract structure. How signals propagate through the ESN and the effects of different weight sets in the network.

Once the weights of the output layer are calculated, the network is ready and the state of the reservoir layer is used to calculate the output of the network as shown in (3).

$$\mathbf{y}(n+1) = f^{\text{out}}\left(\mathbf{W}_{\text{out}}\mathbf{x}(n+1)\right) \tag{3}$$

where $\mathbf{y}(n+1)$ is the output of the network, $\mathbf{W}_{\text{out}}$ is the weight matrix at the readout layer and f^{out} is the readout layer's activation function.

III. ESN in Hardware

Hardware implementations of ESNs are more effective in meeting the critical requirements of applications, such as therapeutic devices and body sensors, in terms of power consumption, processing speed, and area requirements. Current mode circuits, which draw extremely low power, are used for this purpose. From a circuit point of view, an ESN consists of a number of neurons connected by a set of synapses in a specific pattern. Therefore, the primitives required to build an ESN are:

- Architectural topology of the reservoirs
- Input and output processing layers
- Memristive synapse circuit models
- Neuron circuit models

A. Architectural topology of the reservoirs

Topology is defined by the interconnection pattern within the reservoir nodes. Implementing ESN reservoirs with complex topologies incurs a large hardware cost from the standpoint of routing complexity, area overhead, and power consumption. Two hardware friendly topologies were explored in this work, random and ring. Random topology (Fig. 3(b)) is the native pattern of interconnections in which the nodes are randomly connected based on varied degree of connectivity. A dense hardware architecture is needed to implement this type of topology. The second topology is the ring shaped interconnection (Fig. 3(a)), introduced in [9]. In this topology each node is connected to only two neighbors. Due to its low degree of connectivity, it can be easily realized in hardware.

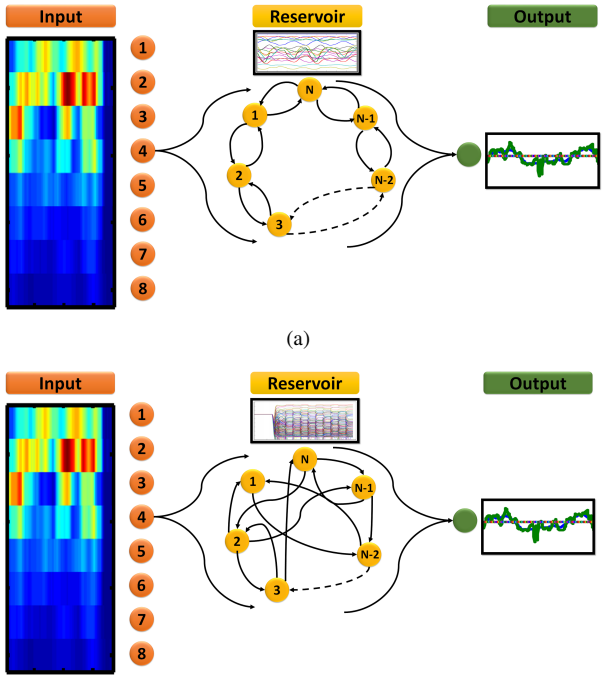

Fig. 3. Block level representations of the two ESN topologies that are explored.(a) Ring topology and (b) Random topology.

B. Memristive Synapse Circuits

A memristor is a non-volatile variable resistor, with state-dependent Ohms Law where its resistance depends on the internal state GAMA [10]. Memristors have been widely used to implement synaptic circuits [11, 12] due to their small footprint, simple device structure, and most importantly zero static power dissipation[13]. Two synaptic circuits were used in this work. The inhibitory synapse (Fig. 4(a)) draws current away from a post-synaptic neuron, similar to GABAergic synapse in the biological brain. The excitatory synapse (Fig. 4(b)) supplies current to the post-synaptic neuron, similar to glutamatergic synapse in biological brain. In both the inhibitory and excitatory synapses, two memristors in parallel divide the input current based on the memristors conductance. Consequently the weight is given as a ratio of conductances as in (4). The output currents inhibit or excite the post-synaptic neuron.

$$w_{-(+)} = \frac{G_{2(4)}}{G_{1(3)} + G_{2(4)}}, \tag{4}$$

where w is the weight of the synapse and $G = 1/R$ is the conductance of memristor in Fig. 4.

C. Neuron Circuits

A current-mode neuron circuit with sigmoid ('s') shaped activation function is used in this design (Fig. 5). This circuit consists of a MOSFET differential pair and a current mirror. The transistor **M2** of the differential pair is grounded which makes the output current $\mathbf{i}_{\text{out}}$ dependent only on the input current $\mathbf{i}_{\text{in}}$. The input resistance Rin can be used to adjust

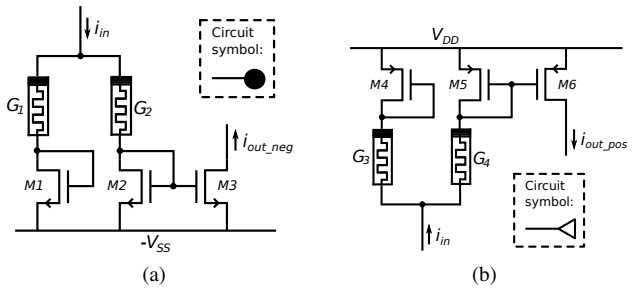

Fig. 4. (a) Inhibitory memristive synapse circuit and (b) excitatory memristive synapse circuit. These circuits are inspired by the function of biological inhibitory (*e.g.* GABAergic) and excitatory (*e.g.* glutamatergic) synapses with ionotropic receptors.

the sigmoid slope. The current flow in the differential pair is mirrored to the output through the transistor **M5**.

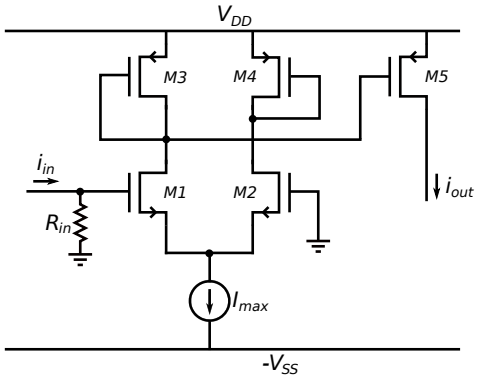

Fig. 5. Neuron circuit for the reservoir and output layers of the ESN with a sigmoid ('s') shaped activation function.

IV. Speech Emotion Recognition

In speech emotion recognition, the emotional status of a human such as anger, fear, happiness etc. are determined based on the speech signals. Human-computer interaction is a classical application of emotion recognition [14]. This property can facilitate better human-computer interaction. Using speech for emotion recognition is simpler and requires less computational resources compared to other inputs such as facial expressions. The Berlin database of Emotional Speech was used for training and learning purposes. The database is publicly available at $http://www.expressive-speech.net/$. In this database, ten actors (five male and five female) recorded 800 utterances. Ten different daily used German sentences were recorded in seven different emotional statuses (anger, joy, sadness, fear, disgust and boredom, and neutral). The utterances were recorded at 16 kHz sampling frequency with 16 bits resolution.

A. Feature Extractor

Selection of features associated with emotions is an important step before feeding inputs into the ESN. These features should be independent of the speaker or lexical content. In speech, emotion is communicated over varying of temporal dynamics of the audio signal [15]. The human ear processes the tone as a non-linear function of the voice frequency. It linearly processes the frequencies below 1000 Hz; but its perception

of frequencies more than 1000 Hz is logarithmic. To extract the emotion content as nearest as the human method, the Mel frequency scale is used [16]. Fig. 6 shows a schematic representation of the feature extractor used in this work. The input audio signal $x(n)$ is divided into shorter pieces of 1600 samples. These pieces are overlapped with 640 samples. The Fast Fourier Transform $\mathbf{X}(k)$ of those vectors are passed to the different Mel-Filter bands. The limits of the eight bands are selected to mimic the human auditory system. The frequency responses of these triangular filters are calculated based on (5).

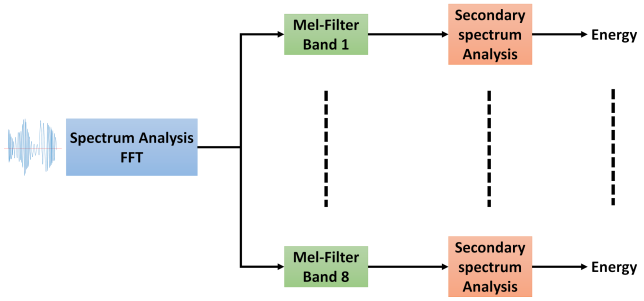

Fig. 6. Block diagram of the feature extractor used in this work. The input is an audio signal and the final output is the energy corresponding to the emotion contents of the input signal.

$$\mathbf{H}_i[k] = \begin{cases} \frac{2(k-bi)}{(d_i-b_i)-(c_i-b_i)}, & \text{if } b_i \le k \le c_i \\ \frac{2(di-k)}{(d_i-b_i)-(d_i-c_i)}, & \text{if } c_i \le k \le d_i \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where i is the index of the filter, $\mathbf{H}_i$ is the response of the ith filter. b_i, c_i and d_i are the start, center and end limits of the ith filter.

The amplitude spectrum of the signals analyzed by the Mel-Filter is calculated based on the responses of the frequency of the filter bands and the absolute value of the FFT $\mathbf{H}(k)$ (Equaion 6). This process ends with eight different signals and each signal is analyzed with FFT to calculate the final features. Based on the length and overlapped values of the input pieces, the extraction frequency is 25Hz.

$$m(i) = \sum_{k=0}^{K} H_i[k]\,|X[k]| \quad (6)$$

Fig. 7 shows the final features of ten neutral and ten anger statuses. It shows that the response of the third-eighth Mel-Filter bands for the neutral status is very low while the anger statues shows high responses at these bands. This difference is the key to the classification.

B. Simulation Methodology

The synapses and the neuron circuits were simulated in HSPICE. The results were analyzed to build an ideal behavioral model of these circuits. Based on these models, the whole system was emulated in MATLAB, for a realistic simulation. Using this approach we were able to analyse the system in detail and find the appropriate parameters to enhance the accuracy. The proposed ESN design was used to classify two emotion statuses (Neutral and Anger). The restricted analysis

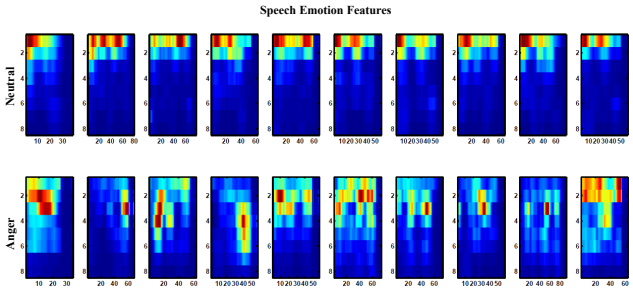

Fig. 7. Ten randomly chosen final emotion features for neutral and anger statuses. The length of each of the signal depends on the length of the actual input audio signal to the feature extractor.

of two emotional states was chosen to simplify hardware testing and reservoir parameters optimization. A total of 156 audio signals were used for both training and testing purposes (110 for training and 46 for testing). The system was tested with different parametric values as explained in the next section.

V. Results and Analysis

Repeated sets of simulations were conducted to find the best values of the size of the reservoir, degree of connectivity and the short memory parameter alpha. 1000 individual simulation runs were conducted with different reservoir sizes (10-500) and degrees of connectivity(5-100%). Fig. 8 shows the testing accuracy of each simulation. It was found that the 190 node reservoir with 20% degree of connectivity has the best accuracy. These values were used to conduct another experiment to find the best Alpha value. The experiment included 100 distinct simulation runs with different Alpha values (0.01-1.0). Fig. 9 shows results of this experiment where the best testing accuracy was achieved when Alpha value ≈0.25.

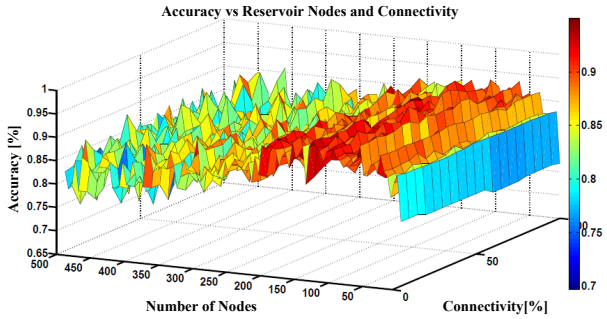

Fig. 8. The effects of the number of nodes within the reservoir and the degree of connectivity of those nodes on the testing accuracy at Alpha ≈0.25 The best accuracy is observed at ≈190 nodes and ≈ 20 % connectivity.

One readout layer node was sufficient to classify the two emotional status (Neutral and Anger). The output of this node is compared against a threshold value to calculate the final binary output (Fig. 10). Different threshold values were used to enhance the classification accuracy. Fig. 11 shows best training and testing accuracy versus different threshold values. Both training and testing reached ≈96% accuracy.

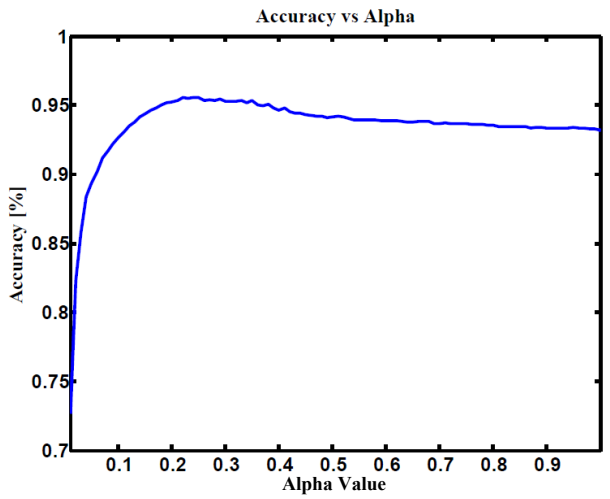

Fig. 9. The short memory parameter Alpha verses testing accuracy at 190 reservoir nodes with 20 % connectivity. Best accuracy is observed at Alpha ≈0.25.

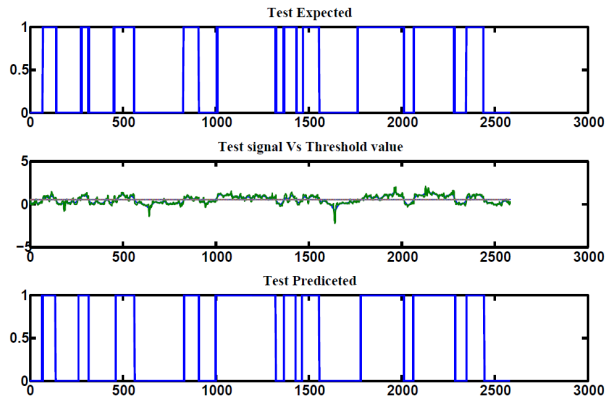

Fig. 10. The expected output of the test case and the actual output of the ESN before and after comparing with the threshold value.

VI. Conclusions

This work introduces a computational architecture of the ESN using memristive circuits for real-time speech emotion recognition. Using only two memristors, the synapses have captured the accurate weight values in the Ring topology.This design successfully recognizes two emotional status (Neutral and Anger) with ≈96% accuracy. A total of 156 audio signals were used for both training and testing purposes.The effect of reservoir learning rate, number of nodes, and degree of connectivity is also studied. Future work will focus on recognizing additional emotional statuses (such as joy, sadness, fear, disgust and boredom) and improving the accuracy by considering the latency in the output signal.

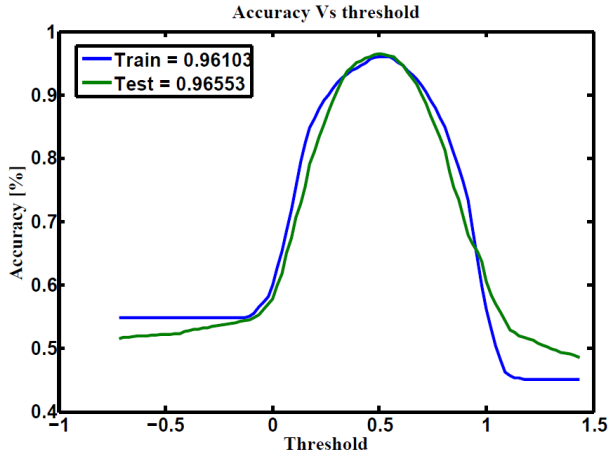

Fig. 11. The classification accuracy of ideal hardware behavior model of the ESN with 190 reservoir nodes at 20% degree of connectivity and Alpha ≈0.25 The best test accuracy is 96.5%.

References

[1] H. Jaeger, "The echo state approach to analysing and training recurrent neural networks – with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, p. 34, 2001.

[2] S. Scherer, M. Oubbati, F. Schwenker, and G. Palm, "Real-time emotion recognition from speech using echo state networks," in *Artificial neural networks in pattern recognition*. Springer, 2008, pp. 205–216.

[3] R. Sacchi, M. C. Ozturk, J. C. Principe, A. A. Carneiro, and I. N. da Silva, "Water inflow forecasting using the echo state network: a brazilian case study," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*. IEEE, 2007, pp. 2403–2408.

[4] M. H. Tong, A. D. Bickett, E. M. Christiansen, and G. W. Cottrell, "Learning grammatical structure with echo state networks," *Neural Networks*, vol. 20, no. 3, pp. 424–432, 2007.

[5] K. Ishu, T. van Der Zant, V. Becanovic, and P. Ploger, "Identification of motion with echo state network," in *OCEANS'04. MTTS/IEEE TECHNO-OCEAN'04*, vol. 3. IEEE, 2004, pp. 1205–1210.

[6] M. D. Skowronski and J. G. Harris, "Automatic speech recognition using a predictive echo state network classifier," *Neural networks*, vol. 20, no. 3, pp. 414–423, 2007.

[7] M. Lukoševičius, H. Jaeger, and B. Schrauwen, "Reservoir computing trends," *KI-Künstliche Intelligenz*, vol. 26, no. 4, pp. 365–371, 2012.

[8] H. Jaeger, "Adaptive nonlinear system identification with echo state networks," in *Advances in neural information processing systems*, 2002, pp. 593–600.

[9] A. Rodan and P. Tino, "Minimum complexity echo state network," *Neural Networks, IEEE Transactions on*, vol. 22, no. 1, pp. 131–144, 2011.

[10] L. Chua, "Resistance switching memories are memristors," *Applied Physics A*, vol. 102, no. 4, pp. 765–783, 2011.

[11] I. E. Ebong and P. Mazumder, "CMOS and memristor-based neural network design for position detection," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 2050–2060, Jun. 2012.

[12] J. Rajendran, H. Manem, R. Karri, and G. S. Rose, "An energy-efficient memristive threshold logic circuit," *IEEE Transactions on Computers*, vol. 61, no. 4, pp. 474–487, 2012.

[13] D. Kuzum, S. Yu, and H. P. Wong, "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, p. 382001, 2013.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[15] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of affective sciences*, pp. 433–456, 2003.

[16] Y. L. X. Jiang, "A kind of audio data retrieval method based on mfcc[j]," *Computer and Digtal Engineering*, vol. 36, no. 9, pp. 19–21, 2008.